Drum Sound Recognition for Polyphonic Audio Signals by Adaptation and Matching of Spectrogram Templates With Harmonic Structure Suppression

Kazuyoshi Yoshii, *Student Member, IEEE*, Masataka Goto, and Hiroshi G. Okuno, *Senior Member, IEEE*

***Abstract*—This paper describes a system that detects onsets of the bass drum, snare drum, and hi-hat cymbals in polyphonic audio signals of popular songs. Our system is based on a template-matching method that uses power spectrograms of drum sounds as templates. This method calculates the distance between a template and each spectrogram segment extracted from a song spectrogram, using Goto's distance measure originally designed to detect the onsets in drums-only signals. However, there are two main problems. The first problem is that appropriate templates are unknown for each song. The second problem is that it is more difficult to detect drum-sound onsets in sound mixtures including various sounds other than drum sounds. To solve these problems, we propose template-adaptation and harmonic-structure-suppression methods. First of all, an initial template of each drum sound, called a seed template, is prepared. The former method adapts it to actual drum-sound spectrograms appearing in the song spectrogram. To make our system robust to the overlapping of harmonic sounds with drum sounds, the latter method suppresses harmonic components in the song spectrogram before the adaptation and matching. Experimental results with 70 popular songs showed that our template-adaptation and harmonic-structure-suppression methods improved the recognition accuracy and achieved 83%, 58%, and 46% in detecting onsets of the bass drum, snare drum, and hi-hat cymbals, respectively.**

***Index Terms*—Drum sound recognition, harmonic structure suppression, polyphonic audio signal, spectrogram template, template adaptation, template matching.**

I. Introduction

THE importance of music content analysis for musical audio signals has been increasing in the field of music information retrieval (MIR). MIR aims at retrieving musical pieces by executing a query about not only text information such as artist names and music titles but also musical contents such as rhythms and melodies. Although the amount of digitally recorded music available over the Internet is rapidly increasing, there are only a few ways of using text information to efficiently find our desired musical pieces in a huge music database. Music content analysis enables MIR systems to automatically understand the contents of musical pieces and to deal with them even if they do not have metadata about the artists and titles.

As the first step of achieving content-based MIR systems in the future, we focus on detecting onset times of individual musical instruments. In this paper, we call this process *recognition*, which means simultaneous processing of both *onset detection* and *identification* of each sound. Although onset time information of each musical instrument is low-level musical content, the recognition results can be used as a basis for higher-level music content analysis concerning the rhythm, melody, and chord, such as beat tracking, melody detection, and chord change detection.

In this paper, we propose a system of *recognizing* drum sounds in polyphonic audio signals sampled from commercial compact-disc (CD) recordings of popular music. We allow various music styles for popular music, such as rock, dance, house, hip-hop, eurobeat, soul, R&B, and folk. Our system *detects onset times* of three drum instruments—bass drum, snare drum, and hi-hat cymbals—while *identifying* them. For a large class of popular music with drum sounds, these three instruments play important roles as the rhythmic backbone of music. We believe that accurate onset detection of drum sounds is useful for describing temporal musical contents such as rhythm, tempo, beat, and measure. Previous studies [1]–[4] on describing those temporal contents, however, have focused on the periodicity of time-frame-based acoustic features, and have not tried to detect accurate onset times of drum sounds. Previous studies [5], [6] on genre classification did not consider onset times of drum sounds while such onset times could be used for improving classification performances by identifying drum patterns unique to musical genres. Some recent studies [7], [8] reported the use of drum patterns for genre classification while Ellis *et al.* [7] dealt with only MIDI signals. The results of our system are useful for such genre classification with higher-level content analysis of real-world audio signals.

The rest of this paper is organized as follows. In Section II, we describe the current state of drum sound recognition techniques. In Section III, we examine the problems and solutions of recognizing drum sounds contained in commercial CD recordings. Sections IV and V describe the proposed solutions: template-adaptation and template-matching methods, respectively. Section VI describes a harmonic-structure-suppression method to improve the performance of our system. Section VII shows experimental results of evaluating these methods. Finally, Section VIII summarizes this paper.

Manuscript received February 1, 2005; revised December 19, 2005. This work was supported in part by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Grant-in-Aid for Scientific Research (A) 15200015 and by the COE Program of MEXT, Japan. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Michael Davies.

K. Yoshii and H. G. Okuno are with the Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan (e-mail: yoshii@kuis.kyoto-u.ac.jp; okuno@i.kyoto-u.ac.jp).

M. Goto is with the National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba 305-8568, Japan (e-mail: m.goto@aist.go.jp).

Digital Object Identifier 10.1109/TASL.2006.876754

1558-7916/$20.00 © 2006 IEEE

II. Art of Drum Sound Recognition

We start on describing the current state of the art of drum sound recognition and related work motivating our approach.

A. Current State

Although there are many studies on onset detection or identification of drum sounds, a few of them have dealt with drum sound recognition for polyphonic audio signals such as commercial CD recordings. The drum sound recognition method by Goto and Muraoka [9] was the earliest work that could deal with drum-sound mixtures of solo performances with MIDI rock-drums. Herrera *et al.* [10] compared conventional feature-based classifiers in the experiments of identifying monophonic drum sounds. To recognize drum sounds in drums-only audio signals, various modeling methods such as N-grams [11], probabilistic models [12], and SVM [13] have been used. By using a noise-space-projection method, Gillet and Richard [14] tried to recognize drum sounds in polyphonic audio signals. These studies, however, cannot fully deal with both the variation of drum-sound features and their distortion caused by the overlapping of other sounds.

The detection of bass and snare drum sounds in polyphonic CD recordings was mentioned in Goto's study on beat tracking [15]. Since it roughly detected them to estimate a hierarchical beat structure, the accurate drum detection was not investigated. Gouyon *et al.* [16] proposed a method that classifies mixed sounds extracted from polyphonic audio signals into two categories of the bass and snare drums. As the former step of the classification, they proposed a percussive onset detection method. It was based on a unique idea of template adaptation that can deal with drum-sound variations according to musical pieces. Zils *et al.* [17] tried the extraction and resynthesis of drum tracks from commercial CD recordings by extending Gouyon's method, and showed the promising results.

To recognize drum sounds in audio signals of drum tracks, sound source separation methods have been focused. They made various assumptions in decomposing a single music spectrogram into multiple spectrograms of musical instruments; independent subspace analysis (ISA) [18], [19] assumes the statistical independence of sources, non-negative matrix factorization (NMF) [20] assumes their non-negativity, and sparse coding combined with NMF [21] assumes their non-negativity and sparseness. Further developments were made by FitzGerald *et al.* [22], [23]. They proposed PSA (Prior Subspace Analysis) [22] that assumes prior frequency characteristics of drum sounds, and applied it to recognize drum sounds in the presence of harmonic sounds [23]. For the same purpose, Dittmar and Uhle [24] adopted non-negative independent component analysis (ICA) that considers the non-negativity of sources. In these studies, the recognition results depend not only on the separation quality but also on the reliability of estimating the number of sources and classifying them. However, the estimation and classification methods are not robust enough for the sake of recognizing drum sounds in audio signals containing time-frequency-varying various sounds.

Klapuri [25] reported a method of detecting onsets of all sounds in polyphonic audio signals. Herrera *et al.* [26] used Klapuri's algorithm to estimate the amount of percussive onsets. However, drum sound identification was not evaluated. To identify drum sounds extracted from polyphonic audio signals, Sandvold *et al.* [27] proposed a method that adapts feature models to those of drum sounds used in each musical piece, but they used correct instrument labels for the adaptation.

B. Related Work

We explain two related methods in detail.

1) Drum Sound Recognition for Solo Drum Performances: Goto and Muraoka [9] reported a template-matching method for recognizing drum sounds contained in musical audio signals of popular-music solo drum performances by a MIDI tone generator. Their method was designed *in the time-frequency domain*. First, a fixed-time-length power spectrogram of each drum to be recognized is prepared as a spectrogram template. There were nine templates corresponding to nine drum instruments (bass and snare drums, toms, and cymbals) in a drum set. Next, onset times are detected by comparing the template with the power spectrogram of the input audio signal, assuming that the input signal is a polyphonic sound mixture of those templates. In the template-matching stage, they proposed a distance measure (we call this "Goto's distance measure" in this paper), which is robust for the spectral overlapping of a drum sound corresponding to the target template with other drum sounds.

Although their method achieved the high recognition accuracy, it has a limitation that the power spectrogram of each drum used in the input audio signal must be registered with the recognition system. In addition, it has difficulty recognizing drum sounds included in polyphonic music because it does not assume the spectral overlapping of harmonic sounds.

2) Drum Sound Resynthesis From CD Recordings: Zils *et al.* [17] reported a template-adaptation method for recognizing bass and snare drum sounds from polyphonic audio signals sampled from popular-music CD recordings. Their method is defined *in the time domain*. First, a fixed-time-length signal of each drum is prepared as a waveform template, which is different from an actual drum signal used in a target musical piece. Next, by calculating the correlation between each template and the musical audio signal, onset times at which the correlation is large are detected. Finally, a drum sound is created (i.e., the signal template is updated) by averaging fixed-time-length signals starting from those detected onset times. These operations are repeated until the template converges.

Although their time-domain analysis seems to be promising, it has limitations in dealing with overlapping drum sounds in the presence of other musical instrument sounds.

III. Drum Sound Recognition Problem for Polyphonic Audio Signals

First, we define the task of our drum sound recognition system. Next, we describe the problems and solutions in recognizing drum sounds in polyphonic audio signals.

A. Target

The purpose of our research is to detect onset times of three kinds of drum instruments in a drum set: bass drum, snare drum, and hi-hat cymbals. Our system takes polyphonic musical audio

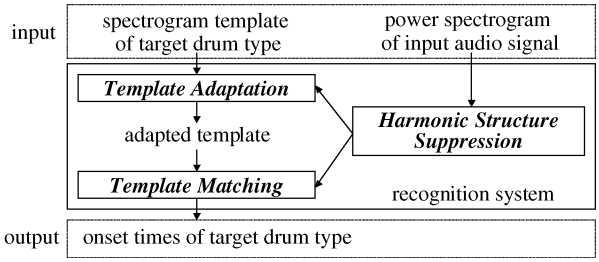

Fig. 1. Overview of drum sound recognition system: a drum-sound spectrogram template (input) is adapted to actual drum-sound spectrograms appearing in the song spectrogram (input) in which the harmonic structure is suppressed. The adapted template is compared with the song spectrogram to detect onsets (output).

signals as input, which are sampled from popular-music CD recordings and contain sounds of vocal parts and various musical instruments (e.g., piano, trumpet, and guitar) as well as drum sounds. Drum sounds are performed by real drum sets (e.g., popular/rock drums) or electronic instruments (e.g., MIDI tone generators). Assuming the main target is popular rock-style music, we focus on the basic playing style of drum performances using normal sticks, and do not deal with special playing styles (e.g., head-mute and brush).

B. Problems

In this paper, we develop a template-based recognition system that defines a template as a fixed-time-length power spectrogram of each drum: bass drum, snare drum, or hi-hat cymbals. There are the following two problems, considering the discussion in Section II-B.

1) Individual Difference Problem: Acoustic features of drum sounds vary among musical pieces and the appropriate templates for recognizing drum sounds in each piece are usually unknown in advance.

2) Mixed Sound Problem: It is difficult to accurately detect drum sounds included in polyphonic audio signals because acoustic features are distorted by the overlapping of other musical instrument sounds.

C. Approach

We propose an advanced template-adaptation method to solve the individual difference problem described in Section III-B. After performing the template adaptation, we detect onset times of drum sounds using an advanced template-matching method. In addition, in order to solve the mixed sound problem, we propose a harmonic-structure-suppression method that improves the robustness of our adaptation and matching methods. Fig. 1 shows an overview of our proposed drum sound recognition system.

1) Template Adaptation: The purpose of this adaptation is to obtain a spectrogram template that is adapted to its corresponding drum sound used in the polyphonic audio signal of a target musical piece. Before the adaptation, we prepare individual spectral templates (we call *seed-templates*) for bass drum, snare drum, and hi-hat cymbals; three templates in total. To adapt the seed-templates to the actual drum sounds, we extended Zils' method to the time-frequency domain.

2) Template Matching: The purpose is to detect all the onset times of drum sounds in the polyphonic audio signal of the target piece, even if other musical instrument sounds overlap the drum sounds. By using Goto's distance measure considering the spectral overlapping, we compare the adapted template with the spectrogram of the audio signal. We present an improved spectral weighting algorithm based on Goto's algorithm for use in calculating the matching distance.

3) Harmonic Structure Suppression: The purpose is to suppress harmonic components of other instrument sounds in the audio signal when recognizing sounds of bass and snare drums. In the recognition of hi-hat cymbal sounds, this processing is not performed under the assumption that harmonic components are weak enough at a high-frequency band.

We use two different distance measures between the template adaptation and matching stages. In the adaptation stage, it is desirable to detect only semi-pure drum sounds that have little overlap with other sounds. Those drum sounds tend to result in a good adapted template that includes little spectral components of other sounds. Because it is not necessary to detect all the onset times of a target drum instrument, a distance measure used in this stage does not care about the spectral overlapping of other sounds. In the matching stage, on the other hand, we used the Goto's distance measure because it is necessary to exhaustively detect all the onset times even if target drum sounds are overlapped by other sounds.

The recognition of bass drum, snare drum, and hi-hat cymbal sounds is performed separately. In the following sections, the term "drum" means one of these three drum instruments.

IV. Template Adaptation

A drum sound template is a power spectrogram in the time-frequency domain. Our template-adaptation method uses a single initial template, called a "*seed template*," for each kind of drum instruments. To recognize the sounds of the bass drum, snare drum and hi-hat cymbals, for example, we require just three seed templates, each of which is individually adapted by using the method.

Our method is based on an iterative adaptation algorithm. An overview of the method is shown in Fig. 2. First, *Onset-Candidate-Detection* stage roughly detects onset candidates in the input audio signal of a musical piece. Starting from each onset candidate, a spectrogram segment whose time-length is fixed is extracted from the power spectrogram of the input audio signal. Then, by using the seed template and all the spectrogram segments, the iterative algorithm successively applies two stages—*Segment Selection* and *Template Updating*—to obtain the adapted template.

1) The *Segment-Selection* stage estimates the reliability that each spectrogram segment includes the drum sound spectrogram. The spectrogram segments with high reliabilities are then selected: this selection is based on a fixed ratio to the number of all the spectrogram segments.
2) The *Template-Updating* stage then reconstructs an updated template by estimating the power that is defined, at each frame and each frequency, as the median power among the selected spectrogram segments. The template is thus adapted to the current piece and used for the next adaptive iteration.

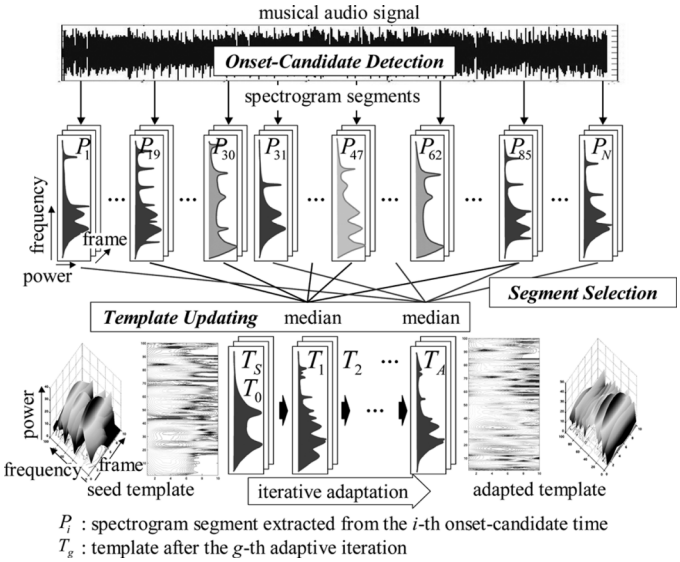

Fig. 2. Overview of template-adaptation method: each template is represented as a fixed-time-length power spectrogram in the time-frequency domain. This method adapts a single seed template corresponding to each drum instrument to actual drum sounds appearing in a target musical piece. The method is based on an iterative adaptation algorithm, which successively applies two stages—*Segment Selection* and *Template Updating*—to obtain the adapted template.

A. Onset Candidate Detection

To reduce the computational cost of the template matching, the *Onset-Candidate-Detection* stage detects possible onset times of drum sounds as candidates: the template matching is performed only at these onset candidates. For the purpose of detecting onset times, Klapuri's method [25] is often used, but we adopted an easy peak-picking method [9] to detect onset *candidate* times. The reason is that it is important to minimize the detection failure (miss) of actual drum-sound onsets; the high recall rate is preferred even if there are many false alarms. Note that each detected onset candidate does not necessarily correspond to an actual drum-sound onset. The template-matching method judges whether each onset candidate is an actual drum-sound onset.

The time at which the power takes a local maximum value is detected as an onset candidate. Let $P(t,f)$ denote the power at frame t and frequency bin f, and $Q(t,f)$ be its time differential. At every frame (441 points), $P(t,f)$ is calculated by applying the short-time Fourier transformation (STFT) with Hanning windows (4096 points) to the signal sampled at 44.1 kHz. In this paper, we use log scale [dB] as the power unit. The onset candidate times are then detected as follows:

1) If $\partial P(t,f)/\partial t > 0$ is satisfied for three consecutive frames $(t = a-1, a, a+1)$, $Q(a,f)$ is defined as

$$Q(a,f) = \left.\frac{\partial P(t,f)}{\partial t}\right|_{t=a}. \tag{1}$$

Otherwise, $Q(a,f) = 0$.

2) At every frame t, the weighted summation $I(t)$ of $Q(t,f)$ is calculated by

$$I(t) = \sum_{f=1}^{2048} F_D(f) Q(t,f) \tag{2}$$

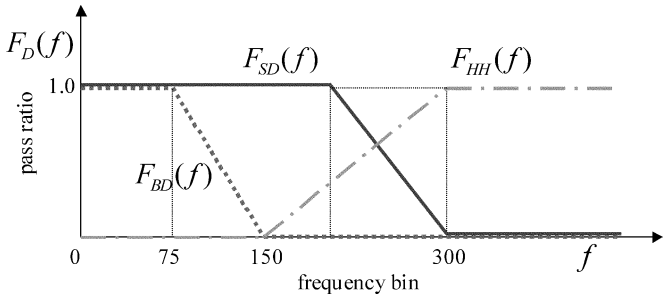

Fig. 3. Lowpass filter functions F_{BD}, F_{SD} which represent typical frequency characteristics of bass and snare drum sounds, and highpass filter function F_{HH} which represents that of hi-hat cymbal sounds.

where $F_D(f)\,\{D|BD, SD, HH\}$ is a lowpass or highpass filter function, as is shown in Fig. 3. We assume that it represents the typical frequency characteristics of bass drum sounds (BD), snare drum sounds (SD) or hi-hat cymbal sounds (HH).

3) Each onset time is given by the time found by peak-picking in $I(t)$. $I(t)$ is smoothed by Savitzky and Golay's smoothing method [28] before its peak time is calculated.

B. Preparing Seed Templates and Spectrogram Segments

1) Seed Template Construction: Seed template T_S (the subscript $_S$ means *seed*) is an average power spectrogram prepared for each drum type to be recognized. The time-length (frames) of seed template T_S is fixed. T_S is represented as a time-frequency matrix whose element is denoted as $T_S(t,f)$ $(1 \le t \le 10$ [frames], $1 \le f \le 2048$ [bins]).

To create seed template T_S, it is necessary to prepare multiple drum sounds each of which contains a solo tone of the drum sound. We used drum-sound samples taken from *"RWC Music Database: Musical Instrument Sound" (RWC-MDB-I-2001)*. They were performed in a normal style on six different real drum sets. By applying the onset candidate detection method, an onset time in each sample is detected. Starting from each time, a power spectrogram whose size is the same as the seed template, is calculated by executing STFT. Therefore, multiple power spectrograms of monophonic drum sounds are obtained, each of which is denoted as S_i $(i = 1, \cdots, N_S)$, where N_S means the number of the extracted power spectrograms (the number of the prepared drum sounds).

Because there are timbre variations of drum sounds, we used multiple drum-sound spectrograms in constructing seed template T_S. Therefore, in this paper, seed template T_S is calculated by collecting the maximum power of the power spectrograms $\{S_i|i = 1, \cdots, N_S\}$ at each frame and each frequency bin

$$T_S(t,f) = \max_{1 \le i \le N_S} S_i(t,f). \tag{3}$$

In the iterative adaptation algorithm, let T_g denote a template being adapted after gth iteration. Because T_S is the first template, T_0 is set to T_S. We also obtain power spectrogram $\acute{T}_g$ weighted by filter function $F_D(f)$

$$\acute{T}_g(t,f) = F_D(f) T_g(t,f). \tag{4}$$

2) Spectrogram Segment Extraction: The ith spectrogram segment P_i $(i = 1, \cdots, N_O)$ is a power spectrogram via STFT starting from an onset candidate time o_i [ms] in the audio signal of a target musical piece, where N_O is the number of the onset candidates. The size of each spectrogram segment P_i is the same with that of seed template T_S, and thus it is also represented as a time-frequency matrix. We also obtain power spectrogram $\acute{P}_i$ weighted by filter function $F_D(f)$

$$\acute{P}_i(t, f) = F_D(f) P_i(t, f). \tag{5}$$

C. Segment Selection

The reliability E_i that spectrogram segment P_i includes the spectral components of the target drum sound is estimated, and then spectrogram segments are selected in descending order with respect to the reliabilities $\{E_i | i = 1, \cdots, N_O\}$. The ratio of the number of the selected segments to the number of all the spectrogram segments (the number of the onset candidates: N_O) is fixed. In this paper, the ratio is empirically set to 0.1 (i.e., the number of the selected segments is $0.1 \times N_O$).

We define the reliability E_i as the reciprocal of the distance D_i between template T_g and spectrogram segment P_i

$$E_i = \frac{1}{D_i}. \tag{6}$$

The distance measure used in calculating the distance D_i is required to satisfy that, if the reliability that spectrogram segment P_i includes the drum sound spectrogram becomes large, the distance D_i becomes small. We describe the individual distance measurement for each drum sound recognition.

1) In Recognition of Bass and Snare Drum Sounds: In the first adaptive iteration, typical spectral distance measures (e.g., Euclidean distance measure) cannot be applied to calculate the distance D_i because those measures inappropriately make the distance D_i large even if spectrogram segment P_i includes the target drum sound spectrogram. In general, the power spectrogram of bass or snare drum sounds has salient spectral peaks that depend on the kind of drum instrument. Because seed template T_0 has never been adapted, the spectral peak positions of T_0 are different from those of the target drum sound spectrogram, which makes the distance D_i large. On the other hand, if spectral peaks of other musical instruments in a spectrogram segment P_i happen to overlap the salient peaks of seed template T_0, the distance D_i becomes small, which results in selecting inappropriate spectrogram segments.

To solve this problem, we perform spectral smoothing in a lower time-frequency resolution for seed template T_0 and each spectrogram segment P_i. In this paper, the time resolution is 2 [frames] and the frequency resolution is 5 [bins] in the spectral smoothing, shown in Fig. 4. This processing allows for differences in the spectral peak positions between seed template T_0 and each spectrogram segment P_i and inhibits the undesirable increase of the distance D_i when a spectrogram segment P_i includes the drum sound spectrogram.

Let $\hat{T}_0$ and $\hat{P}_i$ denote the smoothed seed template and a smoothed spectrogram segment. $\hat{T}_0(t, f)$ in a time-frequency range $2t' - 1 \leq t \leq 2t'$ $(1 \leq t' \leq 5)$, $5f' - 4 \leq f \leq 5f'$ $(1 \leq f' \leq 409)$ is calculated by

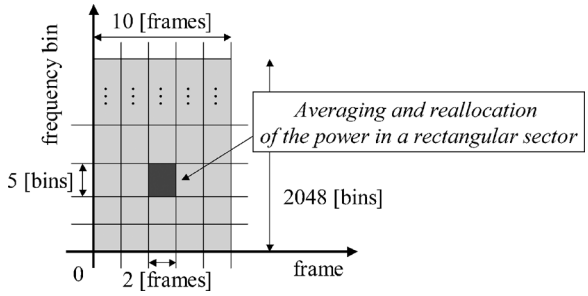

Fig. 4. Spectral smoothing at a lower time-frequency resolution in the *Segment-Selection* stage in bass and snare drum sound recognition: this inhibits the undesirable increase of distance between seed template and spectrogram segment which includes a drum sound spectrogram.

$$\hat{T}_0(t, f) = \frac{1}{10} \sum_{t=2t'-1}^{2t'} \sum_{f=5f'-4}^{5f'} \acute{T}_0(t, f). \tag{7}$$

$\hat{P}_i(t, f)$ is calculated in the same way. This operation means the averaging and reallocation of the power, shown in Fig. 4. First, the time-frequency domain is separated into rectangular sectors. The size of each sector is 2 [frames] $\times$ 5 [bins]. Next, the average power in each sector is calculated, and then reallocated to each bin in that sector.

The spectral distance $D_i^{(0)}$ between seed template T_0 and spectrogram segment P_i in the first iteration is defined as

$$D_i^{(0)} = \sqrt{\sum_{t=1}^{10} \sum_{f=1}^{2048} \left(\hat{T}_0(t, f) - \hat{P}_i(t, f) \right)^2}. \tag{8}$$

After the first iteration, we can use the Euclidean distance measure without the spectral smoothing because the spectral peak positions of template T_g $(g \geq 1)$ are adapted to those of the drum sound used in the audio signal. The spectral distance $D_i^{(g)}$ $(g \geq 1)$ between template T_g and spectrogram segment P_i in the $(g+1)$th adaptive iteration is defined as

$$D_i^{(g)} = \sqrt{\sum_{t=1}^{10} \sum_{f=1}^{2048} \left(\acute{T}_g(t, f) - \acute{P}_i(t, f) \right)^2} \quad (g \geq 1). \tag{9}$$

To focus on the precise characteristic peak positions of the drum sound used in the musical performance, we do not use the spectral smoothing in the equation (9). Because those positions are useful for selecting appropriate spectrogram segments, it is desirable that the equation (9) reflects the differences of the spectral peak positions between the template and a spectrogram segment to the distance.

2) In Recognition of Hi-Hat Cymbal Sounds: The spectral distance D_i in any adaptive iteration is always calculated after the spectral smoothing for template T_g and spectrogram segment P_i. In this paper, the time resolution is 2 [frames] and the frequency resolution is 20 [bins] in the spectral smoothing. A

smoothed template $\hat{T}_g(t,f)$ and a smoothed spectrogram segment $\hat{P}_i(t,f)$ are obtained in the similar way of smoothing the spectrogram of bass and snare drum sounds. Using these spectrograms, the spectral distance D_i between template T_g and spectrogram segment P_i is defined as

$$D_i = \sqrt{\sum_{t=1}^{10}\sum_{f=1}^{2048}\left(\hat{T}_g(t,f) - \hat{P}_i(t,f)\right)^2}. \tag{10}$$

In general, the power spectrogram of hi-hat cymbal sounds seems not to have salient spectral peaks such as those of bass and snare drum sounds. We think it is more appropriate to focus on the shape of the spectral envelope than the fine spectral structure. To ignore the large variation of the local spectral component in a small time-frequency range and extract the spectral envelope, the spectral smoothing is necessary.

D. Template Updating

An updated template is constructed by collecting the median power at each frame and each frequency bin among all the selected spectrogram segments. The updated template is used as the template in the next adaptive iteration. We describe updating algorithms for the template of each drum sound.

1) In Recognition of Bass and Snare Drum Sounds: The updated template $\acute{T}_{g+1}$ which is weighted by filter function F_D is obtained by

$$\acute{T}_{g+1}(t,f) = \underset{1\le i\le N_U}{\text{median}}\, \acute{P}^{(i)}(t,f) \tag{11}$$

where $P^{(i)}$ $(i = 1,\cdots,N_U)$ are the spectrogram segments selected in the *Segment-Selection* stage. N_U is the number of the selected spectrogram segments, which is $0.1 \times N_O$ in this paper.

We pick out the median power at each frame and each frequency bin because we can suppress spectral components that do not belong to the target drum sound spectrogram (Fig. 5). A spectral structure of the target drum sound spectrogram (e.g., salient spectral peaks) can be expected to appear as the same spectral shape in most selected spectrogram segments. On the other hand, spectral components of other musical instrument sounds appear at different frequencies among spectrogram segments. In other words, the local power at the same frame and the same frequency in many spectrogram segments is exposed as the power of the *pure* drum sound spectrogram. By picking out the median of the local power, unnecessary spectral components of other musical instrument sounds become outliers and are not picked out. We can thus obtain a template which is close to the *solo* drum sound spectrogram even if various instrument sounds are included in the musical audio signal.

2) In Recognition of Hi-Hat Cymbal Sounds: The updated and smoothed template $\hat{T}_{g+1}$ that is weighted by filter function F_D is obtained by

$$\hat{T}_{g+1}(t,f) = \underset{1\le i\le N_U}{\text{median}}\, \hat{P}^{(i)}(t,f). \tag{12}$$

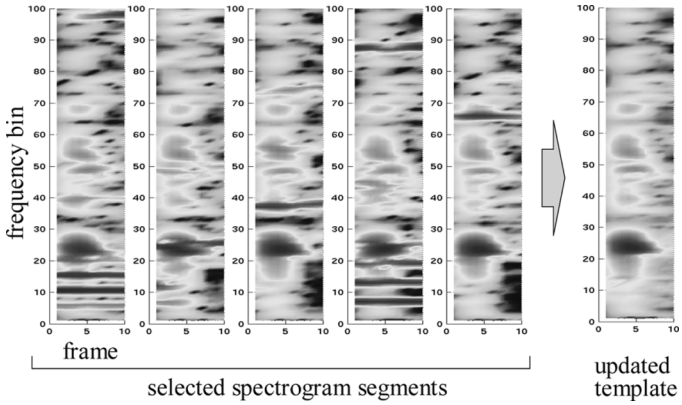

Fig. 5. Updating template by collecting the median power at each frame and each frequency bin among selected spectrogram segments: harmonic components are suppressed in the updated template.

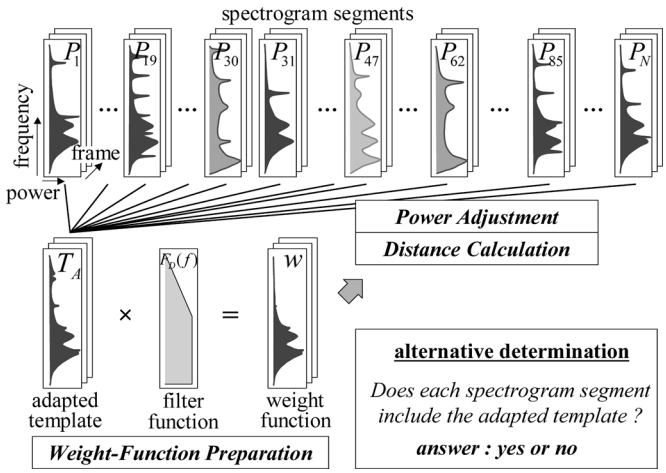

Fig. 6. Overview of template-matching method: each spectrogram segment is compared with the adapted template by using Goto's distance measure to detect actual onset times. This distance measure can appropriately determine whether the adapted template is included in a spectrogram segment even if there are other simultaneous sounds.

If spectrogram segments are not smoothed, the stable median power cannot be obtained because the local power in the spectrogram of hi-hat cymbal sounds varies among onsets. By smoothing the spectrogram segments, the median power is determined as a stable value because the shape of the spectral envelope obtained by the spectral smoothing is stable in the spectrogram of hi-hat cymbal sounds.

V. Template Matching

To find actual onset times, this method judges *whether the drum sound actually occurs at each onset candidate time*, shown in Fig. 6. This alternative determination is difficult because other various sounds often overlap the drum sounds. If we use a general distance measure, the distance between the adapted template and a spectrogram segment including the target drum sound spectrogram becomes large when there are many other sounds that are simultaneously performed with the drum sound. In other words, the overlapping of the other instrument sounds makes the distance large even if the target drum sound spectrogram is included in a spectrogram segment.

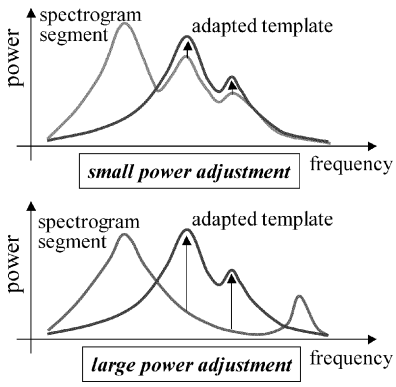

Fig. 7. Power adjustment of spectrogram segments: if a spectrogram segment includes the drum sound spectrogram, the power adjustment value is large (top). Otherwise, the power adjustment value is small (bottom).

To solve this problem, we adopt a distance measure proposed by Goto *et al.* [9]. Because Goto's distance measure focuses on whether the adapted template is included in a spectrogram segment, it can calculate an appropriate distance even if the drum sound is overlapped by other musical instrument sounds. We present an improved method for selecting characteristic frequencies. In addition, we propose a thresholding method that automatically determines appropriate thresholds for each musical piece.

An overview of our method is shown in Fig. 6. First, *Weight-Function-Preparation* stage generates a weight function which represents spectral saliency of each spectral component in the adapted template. This function is used for selecting characteristic frequency bins in the template. Next, *Power-Adjustment* stage calculates the power difference between the template and each spectrogram segment by focusing on the local power difference at each characteristic frequency bin (Fig. 7). If the power difference is larger than a threshold, it judges that the drum sound spectrogram does not appear in that segment, and does not execute the subsequent processing. Otherwise, the power of that segment is adjusted to compensate for the power difference. Finally, *Distance-Calculation* stage calculates the distance between the adapted template and each adjusted spectrogram segment. If the distance is smaller than a threshold, it judges that the drum sound spectrogram is included.

In this section, we describe a template-matching algorithm for bass and snare drum sound recognition. In hi-hat cymbal sound recognition, the adapted template is obtained as the smoothed spectrogram. Therefore, a template-matching algorithm for hi-hat cymbal sound recognition is obtained by replacing [$\acute{}$] with [$\hat{}$] in each expression (e.g., $\acute{T} \rightarrow \hat{T}$, $\acute{P} \rightarrow \hat{P}$).

A. Weight Function Preparation

A weight function represents the spectral saliency at each frame t and frequency bin f in the adapted template. The weight function w is defined as

$$w(t, f) = \acute{T}_A(t, f) \tag{13}$$

where $\acute{T}_A$ represents the adapted template which is weighted by filter function F_D.

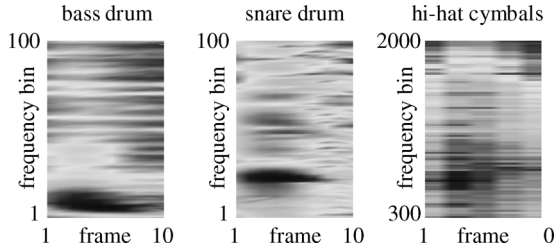

Fig. 8. Examples of adapted templates of bass drum (left), snare drum (center) and hi-hat cymbals (right): these spectrograms show that characteristic frequency bins are different among three drum instruments.

B. Power Adjustment of Spectrogram Segments

The power of each spectrogram segment is adjusted to match with that of the adapted template by assuming that the drum sound spectrogram is included in that spectrogram segment. This adjustment is necessary to correctly determine that the adapted template is included in a spectrogram segment even if the power of the drum sound spectrogram included in that spectrogram segment is smaller than that of the template. On the other hand, if the drum sound spectrogram is not actually included in a spectrogram segment, the power difference is expected to be large. Therefore, if the power difference is larger than a threshold, we determine that the drum sound spectrogram is not included in that spectrogram segment.

To calculate the power difference between each spectrogram segment $\acute{P}_i$ and template $\acute{T}_A$, we focus on the local power differences at spectral characteristic frequency bins of $\acute{T}_A$ in the time-frequency domain. The algorithm of the power adjustment is described as follows:

1) Selecting Characteristic Frequency Bins in Adapted Template: Let $f_{t,k}$ $(k = 1, \cdots, K_D)$ be the characteristic frequency bins in the adapted template, where K_D $(D = BD, SD, HH)$ is the number of characteristic frequency bins at each frame. In this paper, $K_{BD} = 15$, $K_{SD} = 20$, $K_{HH} = 100$. Fig. 8 shows the differences of characteristic frequency bins among three drum instruments. $f_{t,k}$ is determined at each frame t. $f_{t,k}$ is selected as a frequency bin where $w(t, f_{t,k})$ is the kth largest among $w(t, f_{t,k})$ which satisfies the following conditions:

$$w(t, f_{t,k}) \geq w(t, f_{t,k} - 1) \tag{14}$$

$$w(t, f_{t,k}) \geq w(t, f_{t,k} + 1) \tag{15}$$

$$w(t, f_{t,k}) > \lambda_w \times \max_f w(t, f) \tag{16}$$

where λ_w is a constant, which is set to 0.5 in this paper. These three conditions (14), (15), and (16) mean that $w(t, f_{t,k})$ should be peaked along the frequency direction.

2) Calculating Power Difference: The local power difference $\eta_i(t, f_{t,k})$ at frame t and characteristic frequency bin $f_{t,k}$ is calculated as

$$\eta_i(t, f_{t,k}) = \acute{P}_i(t, f_{t,k}) - \acute{T}_A(t, f_{t,k}). \tag{17}$$

The local-time power difference $\delta_i(t)$ at frame t is determined as the first quartile of $\eta_i(t, f_{t,k})$

$$\delta_i(t) = \underset{k}{\text{first-quartile}}\ \eta_i(t, f_{t,k}) \tag{18}$$

$$K_i(t) = \underset{k}{\text{arg-first-quartile}}\ \eta_i(t, f_{t,k}) \tag{19}$$

where $K_i(t)$ is k when $\eta_i(t, f_{t,k})$ is the first quartile. If the number of frames where $\delta_i(t) \leq \Psi_\delta(t)$ is satisfied is larger than a threshold R_δ, we determine that the template is not included in that spectrogram segment, where $\Psi_\delta(t)$ is a threshold automatically determined in Section V-D and R_δ is set to 5 [frames] in this paper.

We pick out not the minimum but the first quartile among the power differences $\{\eta_i(t, f_{t,k})|k = 1, \cdots, K_D\}$ because the latter value is more robust for outliers included in them. The power difference at a characteristic frequency bin may become large when harmonic components of other musical instrument sounds accidentally exist at that frequency. Picking out the first quartile ignores the accidental large power difference and extracts the essential power difference derived from whether the template is included in a spectrogram segment or not.

3) Adjusting Power of Spectrogram Segments: The total power difference Δ_i is calculated by integrating the local-time power difference $\delta_i(t)$ which satisfies $\delta_i(t) > \Psi_\delta(t)$, weighted by weight function w

$$\Delta_i = \frac{\sum_{\{t|\delta_i(t)>\Psi_\delta(t)\}} \delta_i(t) w\left(t, f_{t,K_i(t)}\right)}{\sum_{\{t|\delta_i(t)>\Psi_\delta(t)\}} w\left(t, f_{t,K_i(t)}\right)}. \tag{20}$$

If $\Delta_i \leq \Theta_\Delta$ is satisfied, we are able to determine that the template is not included in that spectrogram segment, where Θ_Δ is a threshold automatically determined in Section V-D.

Let P_i' denote an adjusted spectrogram segment after the power adjustment, obtained by

$$P_i'(t, f) = \acute{P}_i(t, f) - \Delta_i. \tag{21}$$

C. Distance Calculation

To calculate the distance between adapted template $\acute{T}_A$ and an adjusted spectrogram segment P_i', we adopt Goto's distance measure [9]. It is useful for judging whether the adapted template is included in each spectrogram segment or not (the answer is "yes" or "no"). Goto's distance measure does not make the distance large even if the spectral components of the target drum sound are overlapped with those of other sounds. If $P_i'(t, f)$ is larger than $T_A(t, f)$, Goto's distance measure regards $P_i'(t, f)$ as a mixture of spectral components not only of the drum sound but also of other musical instrument sounds. In other words, when we identify that $P_i'(t, f)$ includes $T_A(t, f)$, then the local distance at frame t and frequency bin f is minimized. Therefore, the local distance measure is defined as

$$\gamma_i(t, f) = \begin{cases} 0, & \left(P_i'(t, f) - \acute{T}_A(t, f) \geq \Psi_D\right) \\ 1, & \text{otherwise} \end{cases} \tag{22}$$

where $\gamma_i(t, f)$ is the local distance at frame t and frequency bin f. The negative constant Ψ_D $(D = BD, SD, HH)$ makes this distance measure robust for the small variation of local spectral components. If $P_i'(t, f)$ is larger than about $T_A(t, f)$, $\gamma_i(t, f)$ becomes zero. In this paper, $\Psi_{BD} = \Psi_{SD} = -12.5$ [dB], $\Psi_{HH} = -5$ [dB].

The total distance Γ_i is calculated by integrating the local distance γ_i in the time-frequency domain, weighted by weight function w

$$\Gamma_i = \sum_{t=1}^{10} \sum_{f=1}^{2048} w(t, f) \gamma_i(t, f). \tag{23}$$

To determine whether the targeted drum sound occurred at a time corresponding to the spectrogram segment P_i', the distance Γ_i is compared with a threshold Θ_Γ. If $\Gamma_i < \Theta_\Gamma$ is satisfied, we conclude that the targeted drum sound occurred. Θ_Γ is also automatically determined in Section V-D.

D. Automatic Thresholding

To determine 12 thresholds ($\{\Psi_\delta(t)|t = 1, \cdots, 10\}$, Θ_Δ and Θ_Γ) that are optimized for each musical piece, we use a threshold selection method proposed by Otsu [29]. It is better to dynamically change the thresholds to yield the best recognition results for each piece.

By using Otsu's method, we determine each optimized threshold ($\Psi_\delta(t)$, Θ_Δ or Θ_Γ) which classifies a set of values ($\{\delta_i(t)|i = 1, \cdots, N_O\}$, $\{\Delta_i|i = 1, \cdots, N_O\}$ or $\{\Gamma_i|i = 1, \cdots, N_O\}$) into two classes: the one class contains values which are less than the threshold, the other contains the rest of values. We define a threshold which maximizes the between-class variance (i.e., minimizes the within-class variance).

Finally, to balance the recall rate with the precision rate (these rates are defined in Section VII-A), we adjust thresholds Θ_Δ and Θ_Γ which are determined by Otsu's method

$$\Theta_\Delta \rightarrow \lambda_\Delta \times \Theta_\Delta, \quad \Theta_\Gamma \rightarrow \lambda_\Gamma \times \Theta_\Gamma \tag{24}$$

where λ_Δ and λ_Γ are empirically determined scaling (balancing) factors, which are described in Section VII-B.

VI. Harmonic Structure Suppression

Our proposed method of suppressing harmonic components improves the robustness of the template-adaptation and template-matching methods for the spectral overlapping of harmonic instrument sounds. Real-world CD recordings usually include many harmonic instrument sounds. If the combined power of various harmonic components is much larger than that of the drum sound spectrogram in a spectrogram segment, it is often difficult to correctly detect the drum sound. Therefore, the recognition accuracy is expected to be improved by suppressing those unnecessary harmonic components.

To suppress harmonic components in a musical audio signal, we sequentially perform three operations for each spectrogram segment: estimating F0 of harmonic structure, verifying harmonic components, and suppressing harmonic components. These operations are enabled in bass and snare drum sound recognition. In hi-hat cymbal sound recognition, the harmonic-structure-suppression method is not necessary because most influential harmonic components are expected to be suppressed by the highpass filter function F_{HH}.

A. F0 Estimation of Harmonic Structure

The F0 is estimated at each frame by using a comb-filter-like spectral analysis [30], which is effective in roughly estimating predominant harmonic structures in polyphonic audio signals. The basic idea is to evaluate the reliability $P_{F_0}(t,F)$ that the frequency F is the F0 at each frame t and each frequency F.

The reliability $P_{F_0}(t,F)$ is defined as the summation of the local amplitude weighted by a comb-filter

$$P_{F_0}(t,F)=\sum_{x=0}^{10000} p(x;F)A_i(t,x) \tag{25}$$

where the frequency unit of x and F is [cent],[1] and each increment of x is 100 [cent] in the summation. $A_i(t,x)$ is the local amplitude at frame t and frequency x [cent] in a spectrogram segment P_i. $p(x;F)$ denotes a comb-filter-like function which passes only harmonic components which form the harmonic structure of the F0 F

$$p(x;F)=\sum_{h=1}^{H}\Lambda^{h-1}G(x;F+1200\log_2 h,W_1) \tag{26}$$

$$G(x;m,\sigma)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-m)^2}{2\sigma^2}\right) \tag{27}$$

where H is the number of harmonic components considered and Λ is an amplitude attenuation factor. The spectral spreading of each harmonic component is represented by W_1. $G(x;m,\sigma)$ is a Gaussian distribution, where m is the mean and σ is the standard deviation. In this paper, $H=10$, $A=0.97$, $W_1=150$ [cent].

Frequencies $F_{F_0}(t)$ of the F0 are determined by finding frequencies that satisfy the following condition:

$$P_{F0}\left(t,F_{F_0}(t)\right)>\lambda_{F0}\times\max_{2000\le F\le 7000}P_{F0}(t,F) \tag{28}$$

where λ_{F0} is a constant, which is set to 0.7 in this paper. The F0 is searched from 2000 [cent] (51.9 [Hz]) to 7000 [cent] (932 Hz) by shifting every 100 [cent].

B. Harmonic Component Verification

It is necessary to verify that each harmonic component estimated in Section VI-A is *actually* derived from only harmonic instrument sounds. To suppress all the estimated harmonic components without this verification is not appropriate because a characteristic frequency of drum sounds may be erroneously estimated as a harmonic frequency if the power of drum sounds is much larger than that of harmonic instrument sounds. In another case, a characteristic frequency of drum sounds may be accidentally equal to a harmonic frequency. The verification of each harmonic component prevents characteristic spectral components of drum sounds from being suppressed.

We focus on the general fact that spectral peaks of harmonic components are much more peaked than characteristic spectral peaks of drum sounds. First, the spectral kurtosis $\kappa(h;t,F)$ at frame t in the neighborhood of a hth harmonic component of the F0 F (from $F-50$ [cent] to $F+50$ [cent] in our implementation) is calculated. Second, we determine that the hth harmonic component of the F0 F at frame t is actually derived from only harmonic instrument sounds if $\kappa(h;t,F)$ is larger than a threshold, which is set to 2.0 in this paper (c.f., the kurtosis of the Gaussian distribution is 3.0).

[1]Frequency f_{Hz} in hertz is converted to frequency f_{cent} in cents: $f_{\mathrm{cent}}=1200\log_2(f_{\mathrm{Hz}}/(440\times 2^{(3/12)-5}))$.

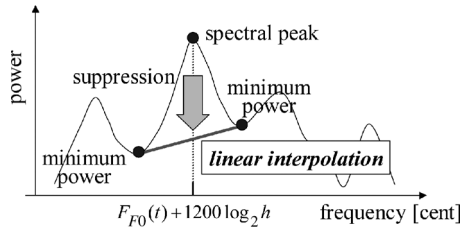

Fig. 9. Suppressing hth harmonic component of the F0 $F_{F0}(t)$ by linearly interpolating between the minimum power on both sides of spectral peak.

C. Harmonic Component Suppression

We suppress harmonic components that are identified as being actually derived from only harmonic instrument sounds. An overview is shown in Fig. 9. First, we find the two frequencies of the local minimum power adjacent to the spectral peak corresponding to each harmonic component at $F+1200\log_2 h$ [cent]. Second, we linearly interpolate the power between them along the frequency axis while preserving the original phase.

VII. Experiments and Results

We performed experiments of recognizing the bass drums, snare drums, and hi-hat cymbals for polyphonic audio signals.

A. Experimental Conditions

We tested our methods on seventy songs sampled from the popular music database *"RWC Music Database: Popular Music" (RWC-MDB-P-2001)* developed by Goto *et al.* [31]. Those songs contain sounds of vocals and various instruments as songs in commercial CDs do. Seed templates were created from solo tones included in *"RWC Music Database: Musical Instrument Sound" (RWC-MDB-I-2001)* [32]: a seed template of each drum is created from multiple sound files each of which contains a sole tone of the drum sound by normal-style performance. All original data were sampled at 44.1 kHz with 16 bits, stereo. We converted them to monaural recordings.

We evaluated the experimental results by the recall rate, precision rate and f-measure

$$\text{recall rate}=\frac{\#(\text{correctly detected onsets})}{\#(\text{actual onsets})}$$

$$\text{precision rate}=\frac{\#(\text{correctly detected onsets})}{\#(\text{detected onsets})}$$

$$\text{f-measure}=\frac{2\cdot\text{recall rate}\cdot\text{precision rate}}{\text{recall rate}+\text{precision rate}}.$$

To prepare actual onset times (correct answers), we extracted onset times (note-on events) of the bass drums, snare drums,

TABLE I
NUMBER OF ACTUAL ONSETS IN 70 MUSICAL PIECES

	bass drums	snare drums	hi-hat cymbals
#(actual onsets)	28084	25393	45219

TABLE II
SETTING OF COMPARATIVE EXPERIMENTS

	template **M**atching (*M*-method)	template **A**daptation (*A*-method)	harmonic **S**uppression (*S*-method)
M-procedure	enabled	-	-
AM-procedure	enabled	enabled	-
SAM-procedure*	enabled	enabled	enabled

**SAM*-procedure is tested in bass or snare drum sound recognition.

and hi-hat cymbals from the standard MIDI files of the seventy songs, which are distributed with the music database, and aligned them to the corresponding audio signals by hand. The number of actual onsets of each drum sound included in seventy songs is shown in Table I (about 100 000 onsets in total). If the difference between a detected onset time and an actual onset time was less than 25 [ms], we judged that the detected onset time is correct.

B. Experimental Results

To evaluate our proposed three methods: template-matching method (M-method), template-adaptation method (A-method), and harmonic-structure-suppression method (S-method), we performed comparative experiments by enabling each method one by one: we tested three procedures shown in Table II, M-procedure, AM-procedure, and SAM-procedure. The SAM-procedure was not tested for recognizing hi-hat cymbal sounds because the S-method is enabled only for recognizing bass or snare drum sounds. The M-procedure used a seed template instead of the adapted template for the template-matching. The balancing factors λ_{Δ} and λ_{Γ} were determined for each experiment as shown in Table III.

For convenience, we evaluated three procedures by dividing 70 musical pieces into three groups: group I, II, and III. First, 70 pieces were sorted in descending order with respect to the f-measure by the fully-enabled procedure (i.e., SAM-procedure in bass and snare drum sound recognition, AM-procedure in hi-hat cymbal sound recognition). Second, the first 20 pieces were put in group I, and the next 25 pieces were put in group II, and the remaining 25 pieces were put in group III.

The average recall and precision rates of onset *candidate* detection was 88%/22% (bass drum sound recognition), 77%/18% (snare drum sound recognition), and 87%/36% (hi-hat cymbal sound recognition). This means the chance rates of onset detection by the coin-toss decision were 29%, 25%, and 39%, respectively. Table III shows the experimental results obtained by each procedure. Table IV shows the recognition error reduction rates which represent the f-measure improvement obtained by enabling the A-method added to the M-procedure, and that obtained by enabling the S-method added to the AM-procedure. Table V shows a complete list of musical pieces sorted in descending order with respect to f-measure of each drum instrument recognition. Fig. 10 shows f-measure curves along the sorted musical pieces in recognizing each drum instrument.

C. Discussion

The experimental results show the effectiveness of our methods. In general, the fully-enabled SAM-procedures yielded the best performance in bass and snare drum sound recognition. In these case, the average f-measure was 82.924% and 58.288%, respectively. In hi-hat cymbal sound recognition by the AM-procedure, the average f-measure was 46.249%. In total, the f-measure averaged over those three drum instruments was about 62%. In our observation, the effectiveness of the A-method and S-method was almost independent to specific playing styles. If harmonic sounds which mainly distribute in a low frequency band (e.g., spectral components of bass line) are more dominant, the suppression method tends to be more effective. We discuss in detail in the following sections.

1) Bass Drum Sound Recognition: The f-measure in bass drum sound recognition (82.92% in total) was highest among the results of recognizing three drum instruments. Table IV showed that both the A-method and the S-method were very effective, especially in group I. It also showed that the S-method in recognizing bass drum sounds was more effective, compared to snare drum sound recognition. The S-method could suppress undesirable harmonic components of the bass line which has the large power in a low frequency band.

2) Snare Drum Sound Recognition: In group I, the f-measure was drastically improved from 65.33% to 87.63% by enabling both the A-method and the S-method. Table IV showed that the S-method in recognizing snare drum sounds was less effective than the A-method.

In group II, on the other hand, the S-method was more effective than the A-method. These results suggest that the template-adaptation became to work correctly after suppressing harmonic components in some pieces. In other words, the A-method and the S-method helped each other in improving the f-measure, and thus it is important to use both methods.

In group III, however, the f-measure was slightly degraded by enabling the A-method because the template-adaptation failed in some pieces. In these pieces, the seed template was erroneously adapted to harmonic components. The S-method was not effective enough to recover from such erroneous adaptation. These facts suggest that acoustic features of snare drum sounds in these pieces are too different from those of the seed template. To overcome these problems, we plan to incorporate multiple templates for each drum instrument.

3) Hi-Hat Cymbal Sound Recognition: The f-measure in hi-hat cymbal sound recognition (46.25% in total) was lowest among the experimental results in recognizing three drum instruments. The performance without the A-method and the S-method indicates that this is the most difficult task in our experiments. Unfortunately, the A-method was not effective enough for hi-hat cymbals, while it reduced some errors as shown in Table IV. This is because there are three major playing styles for hi-hat cymbals, closed, open, and half-open, and they are used in a mixed way in an actual musical piece. Since our method used just a single template, the template could not cover all spectral variations by those playing styles and was not appropriately adapted to those sounds in the piece even by the A-method. We plan to incorporate multiple templates

TABLE III
DRUM SOUND RECOGNITION RATES

		factor		group I	group II	group III	average rate in total		
		λ_Δ	λ_Γ	f-measure	f-measure	f-measure	recall	precision	f-measure
bass drums	*M*-procedure	1.0	1.0	83.45 %	71.01 %	59.75 %	70.132 %	70.295 %	70.214 %
	AM-procedure	1.1	1.5	92.78 %	81.32 %	62.70 %	77.047 %	77.798 %	77.421 %
	SAM-procedure	1.1	1.5	**97.43 %**	**88.81 %**	**66.77 %**	81.235 %	84.684 %	**82.924 %**
snare drums	*M*-procedure	1.1	1.1	65.33 %	58.28 %	**40.15 %**	55.318 %	49.883 %	52.460 %
	AM-procedure	1.2	1.5	82.96 %	58.45 %	36.87 %	58.886 %	51.277 %	54.819 %
	SAM-procedure	1.2	1.5	**87.63 %**	**64.03 %**	36.98 %	55.677 %	61.156 %	**58.288 %**
hi-hat cymbals	*M*-procedure	1.3	1.2	55.78 %	45.22 %	**34.39 %**	47.533 %	42.433 %	44.838 %
	AM-procedure	1.7	1.7	**59.30 %**	**46.09 %**	34.15 %	47.944 %	44.669 %	**46.249 %**

Note: 70 musical pieces were sorted in descending order with respect to the f-measure by the fully-enabled procedure (i.e., SAM-procedure in bass and snare drum sound recognition, AM-procedure in hi-hat cymbal sound recognition). The first 20 pieces were put in group I, and the next 25 ones were put in group II, and the last 25 ones were put in group III.

TABLE IV
RECOGNITION ERROR REDUCTION RATES

		group I	group II	group III	total
bass drums	enabling *A*-method	56.37 %	35.56 %	7.329 %	24.196 %
	enabling *S*-method	64.40 %	40.10 %	10.91 %	24.372 %
snare drums	enabling *A*-method	50.85 %	0.4075 %	-5.556 %	4.9621 %
	enabling *S*-method	27.40 %	13.43 %	0.1724 %	7.6780 %
hi-hat cymbals	enabling *A*-method	7.960 %	1.588 %	-0.1095 %	3.8841 %

Note: The definition of group I, II and III is described in Table III. This shows the recognition error reduction rates which represent the f-measure improvement obtained by enabling the A-method added to the M-procedure, and that obtained by enabling the S-method added to the AM-procedure.

TABLE V
LIST OF MUSICAL PIECES SORTED IN DESCENDING ORDER WITH RESPECT TO F-MEASURE

bass drums	11, 14, 52, 42, 4, 13, 57, 39, 62, 7, 24, 25, 12, 30, 86, 8, 32, 20, 87, 46, 96, 47, 50, 61, 81, 85, 68, 27, 89, 1, 43, 33, 16, 29, 51, 23, 94, 19, 17, 15, 90, 2, 63, 44, 6, 21, 67, 28, 35, 45, 69, 83, 65, 5, 97, 92, 93, 66, 37, 40, 98, 53, 84, 10, 56, 26, 22, 54, 36, 82
snare drums	26, 11, 61, 25, 50, 52, 22, 53, 6, 18, 51, 62, 7, 21, 20, 63, 90, 8, 83, 37, 85, 13, 68, 47, 35, 98, 44, 89, 88, 43, 84, 14, 33, 5, 46, 23, 36, 30, 28, 12, 54, 92, 40, 86, 32, 1, 87, 94, 56, 45, 95, 66, 97, 96, 82, 24, 39, 16, 17, 93, 9, 42, 10, 3, 31, 15, 49, 29, 57, 100
hi-hat cymbals	4, 25, 53, 87, 11, 52, 69, 7, 48, 5, 13, 17, 62, 32, 56, 21, 96, 14, 47, 46, 16, 50, 97, 24, 42, 12, 8, 89, 30, 63, 1, 98, 31, 51, 90, 23, 91, 22, 45, 49, 66, 68, 20, 19, 43, 70, 92, 40, 61, 10, 15, 86, 37, 6, 93, 33, 83, 95, 94, 65, 54, 85, 44, 84, 82, 67, 81, 36, 39, 88

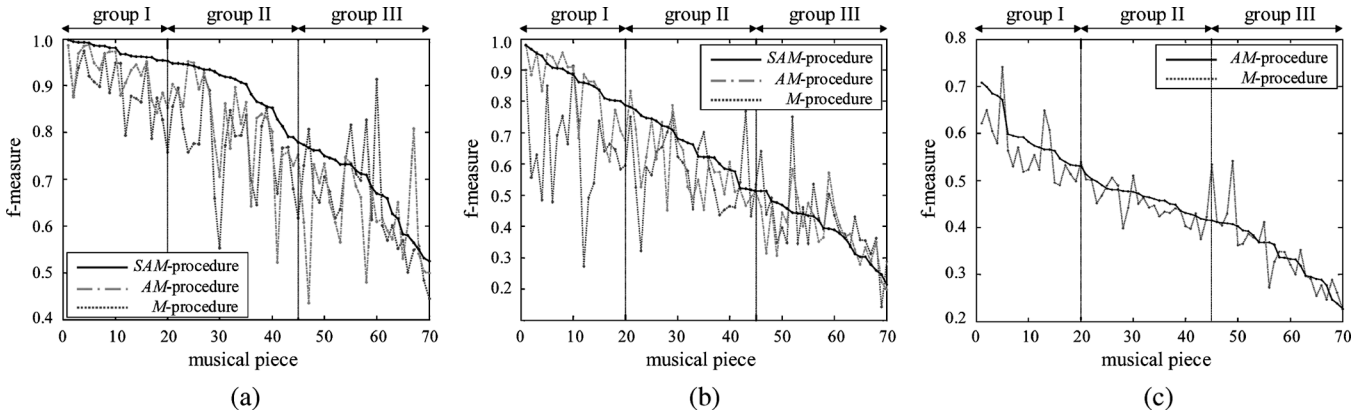

Fig. 10. (a), (b): f-measure curves by three procedures in (a) bass drum sound recognition and (b) snare drum sound recognition along sorted musical pieces in descending order with respect to f-measure by SAM-procedure. (c): f-measure curves by two procedures in hi-hat cymbal sound recognition along sorted musical pieces in descending order with respect to f-measure by AM-procedure.

as discussed above to deal with this difficulty while another problem of identifying the playing styles of hi-hat cymbals will still remain an open question.

VIII. CONCLUSION

In this paper, we have presented a drum sound recognition system that can detect onset times of drum sounds and identify them. Our system used template-adaptation and template-matching methods to individually detect onset times of three drum instruments, the bass drum, snare drum, and hi-hat cymbals. Since a drum-sound spectrogram prepared as a *seed template* is different from one used in a musical piece, our template-adaptation method adapts the template to the piece. By using the adapted template, our template-matching method then detects their onset times even if drum sounds are overlapped by other musical instrument sounds. In addition, to improve the performance of the adaptation and matching, we proposed a harmonic-structure-suppression method that suppresses harmonic components of other musical instrument sounds by using comb-filter-like spectral analysis.

To evaluate our system, we performed reliable experiments with popular-music CD recordings, which are the largest experiments for drum sounds as far as we know. The experimental results showed that both of the template-adaptation and harmonic-structure-suppression methods improved the f-measure of recognizing each drum. The average f-measures were 82.924%, 58.288%, and 46.249% in recognizing bass drum sounds, snare drum sounds, and hi-hat cymbal sounds, respectively. Our system, called *AdaMast* [33], in which the harmonic-structure-suppression method was disabled won the first prize of Audio Drum Detection Contest in MIREX2005. We expect that these results could be used as a benchmark.

In the future, we plan to use multiple seed templates for each kind of the drums to improve the coverage of the timbre variation of drum sounds. A study on timbre variation of drum sounds [34] seems to be helpful. The improvement of the template-matching method is also necessary to deal with the spectral variation among onsets. In addition, we will apply our system to rhythm-related content description for building a content-based MIR system.

REFERENCES

[1] E. Scheirer, "Tempo and beat analysis of acoustic musical signals," *J. Acoust. Soc. Am.*, vol. 103, no. 1, pp. 588–601, Jan. 1998.

[2] J. Paulus and A. Klapuri, "Measuring the similarity of rhythmic patterns," in *Proc. Int. Conf. Music Information Retrieval (ISMIR)*, 2002, pp. 150–156.

[3] F. Gouyon and P. Herrera, "Determination of the meter of musical audio signals: seeking recurrences in beat segment descriptors," in *Proc. Audio Engineering Soc. (AES), 114th Conv.*, 2003.

[4] E. Pampalk, S. Dixon, and G. Widmer, "Exploring music collections by browsing different views," *J. Comput. Music J.*, vol. 28, no. 2, pp. 49–62, summer 2004.

[5] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, Jul. 2002.

[6] S. Dixon, E. Pampalk, and G. Widmer, "Classification of dance music by periodicity patterns," in *Proc. Int. Conf. Music Information Retrieval (ISMIR)*, 2003, pp. 159–165.

[7] D. Ellis and J. Arroyo, "Eigenrhythms: Drum pattern basis sets for classification and generation," in *Proc. Int. Conf. Music Information Retrieval (ISMIR)*, 2004, pp. 554–559.

[8] C. Uhle and C. Dittmar, "Drum pattern based genre classification of popular music," in *Proc. Int. Conf. Audio Eng. Soc. (AES)*, 2004.

[9] M. Goto and Y. Muraoka, "A sound source separation system for percussion instruments," *IEICE Trans. D-II*, vol. J77-D-II, no. 5, pp. 901–911, May 1994.

[10] P. Herrera, A. Yeterian, and F. Gouyon, "Automatic classification of drum sounds: a comparison of feature selection methods and classification techniques," in *Proc. Int. Conf. Music and Artificial Intelligence (ICMAI), LNAI2445*, 2002, pp. 69–80.

[11] J. Paulus and A. Klapuri, "Conventional and periodic N-grams in the transcription of drum sequences," in *Proc. Int. Conf. Multimedia and Expo (ICME)*, 2003, pp. 737–740.

[12] ——, "Model-based event labeling in the transcription of percussive audio signals," in *Proc. Int. Conf. Digital Audio Effects (DAFX)*, 2003, pp. 73–77.

[13] O. Gillet and G. Richard, "Automatic transcription of drum loops," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2004, pp. 269–272.

[14] ——, "Drum track transcription of polyphonic music using noise subspace projection," in *Proc. Int. Conf. Music Information Retrieval (ISMIR)*, 2005.

[15] M. Goto, "An audio-based real-time beat tracking system for music with or without drum-sounds," *J. New Music Res.*, vol. 30, no. 2, pp. 159–171, Jun. 2001.

[16] F. Gouyon, F. Pachet, and O. Delerue, "On the use of zero-crossing rate for an application of classification of percussive sounds," in *Proc. COST-G6 Conf. Digital Audio Effects (DAFX)*, 2000.

[17] A. Zils, F. Pachet, O. Delerue, and F. Gouyon, "Automatic extraction of drum tracks from polyphonic music signals," in *Proc. Int. Conf. Web Delivering of Music (WEDELMUSIC)*, 2002, pp. 179–183.

[18] D. FitzGerald, E. Coyle, and B. Lawlor, "Sub-band independent subspace analysis for drum transcription," in *Proc. Int. Conf. Digital Audio Effects (DAFX)*, 2002, pp. 65–69.

[19] C. Uhle, C. Dittmar, and T. Sporer, "Extraction of drum tracks from polyphonic music using independent subspace analysis," in *Proc. Int. Symp. Independent Component Analysis and Blind Signal Separation (ICA)*, 2003, pp. 843–848.

[20] J. Paulus and A. Klapuri, "Drum transcription with non-negative spectrogram factorisation," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2005.

[21] T. Virtanen, "Sound source separation using sparse coding with temporal continuity objective," in *Proc. Int. Computer Music Conf. (ICMC)*, 2003, pp. 231–234.

[22] D. FitzGerald, B. Lawlor, and E. Coyle, "Prior subspace analysis for drum transcription," in *Proc. Audio Eng. Soc. (AES), 114th Conv.*, 2003.

[23] ——, "Drum transcription in the presence of pitched instruments using prior subspace analysis," in *Proc. Irish Signals Syst. Conf. (ISSC)*, 2003, pp. 202–206.

[24] C. Dittmar and C. Uhle, "Further steps towards drum transcription of polyphonic music," in *Proc. Audio Eng. Soc. (AES), 116th Conv.*, 2004.

[25] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1999, pp. 3089–3092.

[26] P. Herrera, V. Sandvold, and F. Gouyon, "Percussion-related semantic descriptors of music audio files," in *Proc. Int. Conf. Audio Eng. Soc. (AES)*, 2004.

[27] V. Sandvold, F. Gouyon, and P. Herrera, "Percussion classification in polyphonic audio recordings using localized sound models," in *Proc. Int. Conf. Music Information Retrieval (ISMIR)*, 2004, pp. 537–540.

[28] A. Savitzky and M. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *J. Anal. Chem.*, vol. 36, no. 8, pp. 1627–1639, Jul. 1964.

[29] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-6, no. 1, pp. 62–66, Jan. 1979.

[30] M. Goto, K. Itou, and S. Hayamizu, "A real-time filled pause detection system for spontaneous speech recognition," in *Proc. Eurospeech*, 1999, pp. 227–230.

[31] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: popular, classical, and jazz music databases," in *Proc. Int. Conf. Music Information Retrieval (ISMIR)*, 2002, pp. 287–288.

[32] ——, "RWC music database: music genre database and musical instrument sound database," in *Proc. Int. Conf. Music Information Retrieval (ISMIR)*, 2003, pp. 229–230.

[33] K. Yoshii, M. Goto, and H. Okuno, "AdaMast: a drum sound recognizer based on adaptation and matching of spectrogram templates," in *Proc. Music Information Retrieval Evaluation eXchange (MIREX)*, 2005.

[34] E. Pampalk, P. Hlavac, and P. Herrera, "Hierarchical organization and visualization of drum sample libraries," in *Proc. Int. Conf. Digital Audio Effects (DAFX)*, 2004, pp. 378–383.

Kazuyoshi Yoshii (S'05) received the B.S. and M.S. degrees from Kyoto University, Kyoto, Japan, in 2003 and 2005, respectively. He is currently pursuing the Ph.D degree in the Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University.

His research interests include music scene analysis and human-machine interaction.

Mr. Yoshii is a member of the Information Processing Society of Japan (IPSJ) and Institute of Electronics, Information, and Communication Engineers (IEICE). He is supported by the JSPS Research Fellowships for Young Scientists (DC1). He has received several awards including the FIT2004 Paper Award and the Best in Class Award of MIREX2005.

Masataka Goto received the Doctor of Engineering degree in electronics, information, and communication engineering from Waseda University, Tokyo, Japan, in 1998.

He then joined the Electrotechnical Laboratory (ETL; reorganized as the National Institute of Advanced Industrial Science and Technology (AIST) in 2001), where he has been a Senior Research Scientist since 2005. He served concurrently as a Researcher in Precursory Research for Embryonic Science and Technology (PRESTO), Japan Science and Technology Corporation (JST), from 2000 to 2003, and an Associate Professor in the Department of Intelligent Interaction Technologies, Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba, Japan, since 2005. His research interests include music information processing and spoken language processing.

Dr. Goto is a member of the Information Processing Society of Japan (IPSJ), Acoustical Society of Japan (ASJ), Japanese Society for Music Perception and Cognition (JSMPC), Institute of Electronics, Information, and Communication Engineers (IEICE), and International Speech Communication Association (ISCA). He has received 18 awards, including the IPSJ Best Paper Award and IPSJ Yamashita SIG Research Awards (special interest group on music and computer, and spoken language processing) from the IPSJ, the Awaya Prize for Outstanding Presentation and Award for Outstanding Poster Presentation from the ASJ, Award for Best Presentation from the JSMPC, Best Paper Award for Young Researchers from the Kansai-Section Joint Convention of Institutes of Electrical Engineering, WISS 2000 Best Paper Award and Best Presentation Award, and Interaction 2003 Best Paper Award.

Hiroshi G. Okuno (SM'06) received the B.A. and Ph.D degrees from the University of Tokyo, Tokyo, Japan, in 1972 and 1996, respectively.

He worked for Nippon Telegraph and Telephone, Kitano Symbiotic Systems Project, and Tokyo University of Science. He is currently a Professor in the Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Kyoto, Japan. He was a Visiting Scholar at Stanford University, Stanford, CA, and Visiting Associate Professor at the University of Tokyo. He has done research in programming languages, parallel processing, and reasoning mechanisms in AI, and is currently engaged in computational auditory scene analysis, music scene analysis, and robot audition. He edited (with D. Rosenthal) *Computational Auditory Scene Analysis* (Princeton, NJ: Lawrence Erlbaum, 1998) and (with T. Yuasa) *Advanced Lisp Technology* (London, U.K.: Taylor &Francis, 2002).

Dr. Okuno has received various awards including the 1990 Best Paper Award of JSAI, the Best Paper Award of IEA/AIE-2001 and 2005, and IEEE/RSJ Nakamura Award for IROS-2001 Best Paper Nomination Finalist. He was also awarded 2003 Funai Information Science Achievement Award. He is a member of the IPSJ, JSAI, JSSST, JSCS, RSJ, ACM, AAAI, ASA, and ISCA.